STAT 306 Project

Sophia Yang (33176769),

13 April, 2023

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.2.1     v stringr 1.5.0
## v tidyr   1.3.0     v forcats 0.5.2
## v readr   2.1.3
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(RColorBrewer)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(cowplot)
library(mltools)
```

```
##
## Attaching package: 'mltools'
##
## The following object is masked from 'package:tidyr':
##
##     replace_na
```

```
library(leaps)
```

Introduction

We will be using the sleep efficiency dataset:

https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency in Kaggle, which contains 100 observations.

The data was collected in 2021 by a research team in the UK, and was collected from the University of Oxfordshire. It was collected from a local community over a period of several months using a combination of self-reported surveys, actigraphy, and polysomnography (a sleep monitoring technique).

Our research question and motivation behind the analysis of the data:

Which variables are most important in predicting sleep efficiency? And how do these variables relate to sleep efficiency?

We use a forward selection process to determine which variables are most relevant for predicting sleep efficiency and build regression models based only on the variables selected. After comparing different models, we select the best model and use this model to predict the test dataset and combine it with the actual values to see how accurate our model is.

From the model, we can develop a general idea about what factors relate to sleep efficiency and how they are correlated. Therefore, it may suggest some methods to improve sleep patterns by controlling certain factors.

```
sleep <- read.csv('Sleep_Efficiency.csv')

# Convert certain variables into categorical variables
# Gender: male is encoded as 1, and female is encoded as 0
# Smoking status: Yes is encoded as 1, and No is encoded as 0
sleep <- sleep |>
    mutate(Caffeine.consumption = as.factor(Caffeine.consumption),
           Awakenings = as.factor(Awakenings),
           Alcohol.consumption = as.factor(Alcohol.consumption),
           Smoking.status = as.factor(case_when(
               Smoking.status == 'Yes' ~ 1,
               Smoking.status == 'No' ~ 0,
               TRUE ~ NA)),
            Gender= as.factor(case_when(
               Gender == 'Male' ~ 1,
               Gender == 'Female' ~ 0,
               TRUE ~ NA)),
            Exercise.frequency = as.factor(Exercise.frequency)
           )
```

We include 12 variables:

The response variable:

Sleep efficiency: a numerical variable ranging from 0 to 1 that indicates the proportion of time in bed spent asleep

The explanatory variables are:

1. Age: numerical variable ranges from 9 to 69
2. Gender: categorical / dummy variable with 50% of the data is male and the rest half female

3. Sleep duration: numerical variable indicating the total amount of time the test subject slept (in hours).
4. Awakenings: categorical variable indicating the number of times the test subject wakes up during the night
5. REM sleep percentage: numerical variable indicating the percentage of total sleep time spent in REM (rapid eye movement) sleep
    - REM sleep is the stage where people have intense brain activities (dreams) and restores the areas of the brain that help with memory and learning. 20% of the total sleep time in the REM stage is considered good
6. Deep sleep percentage: numerical variable indicating the percentage of total sleep time spent in Deep (non - rapid eye movement) sleep
    - Deep sleep is important for body to replenish energy stores and repair muscles, bones, and tissue. 15% to 25% of the total sleep time in the deep sleep stage is considered normal
7. Light sleep percentage: numerical variable indicating the percentage of total sleep time spent in Light sleep
    - Light sleep is the transitional stage between waking and sleeping. Typically take up about 50% to 60% or more of the total sleep time
8. Caffeine consumption: categorical variable indicating the amount of caffeine consumed in the 24 hours prior to bedtime (in mg)
9. Alcohol consumption: categorical variable indicating the amount of alcohol consumed in the 24 hours prior to bedtime (in oz)
10. Smoking status: categorical / dummy variable that states whether or not the person smokes
11. Exercise frequency: categorical variable indicating the number of times the person exercises each week

Although some explanatory variables seem to be numerical (e.g., Alcohol consumption, Caffeine consumption), the number of values that these variables take is limited; therefore, we can view these variables as categorical with a few levels.

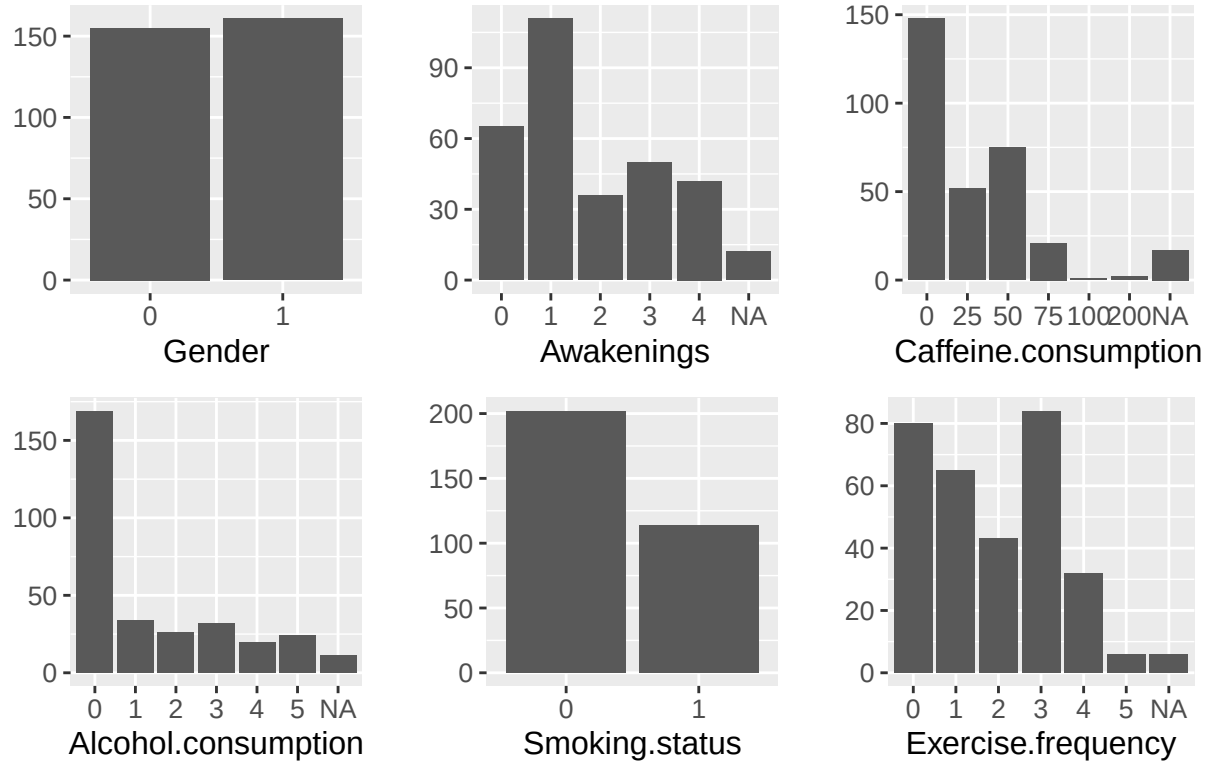
```
sleep_data <- sleep |>
    select(-ID, -Bedtime, -Wakeup.time)
```

```
set.seed(123)
train_ind <- sample.int(nrow(sleep_data), size = nrow(sleep_data) * 0.70, replace = F)
sleep_train <- sleep_data[train_ind,]
sleep_test <- sleep_data[-train_ind,]
```

```
options(repr.plot.height = 15, repr.plot.width = 20)
gender_dist <- sleep_train |>
    ggplot(aes(x = Gender)) +
    geom_bar() +
    ylab('') +
    theme(text = element_text(size = 12))

awaken_dist <- sleep_train |>
    ggplot(aes(x = Awakenings)) +
    geom_bar() +
    ylab('') +
    theme(text = element_text(size = 12))
```

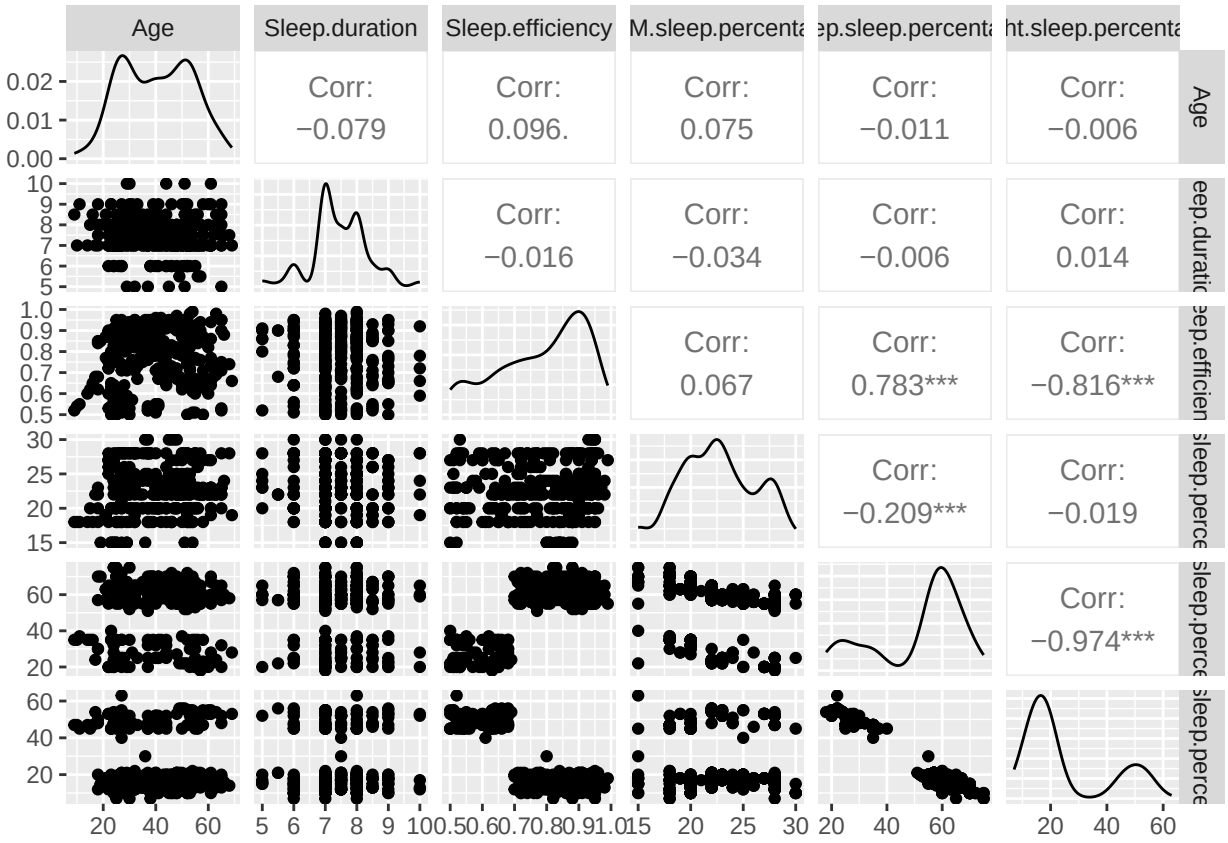
```r
coffee_dist <- sleep_train |>
    ggplot(aes(x = Caffeine.consumption)) +
    geom_bar() +
    ylab('') +
    theme(text = element_text(size = 12))

alcohol_dist <- sleep_train |>
    ggplot(aes(x = Alcohol.consumption)) +
    geom_bar() +
    ylab('') +
    theme(text = element_text(size = 12))

smoke_dist <- sleep_train |>
    ggplot(aes(x = Smoking.status)) +
    geom_bar() +
    ylab('') +
    theme(text = element_text(size = 12))

exercise_dist <- sleep_train |>
    ggplot(aes(x = Exercise.frequency)) +
    geom_bar() +
    ylab('') +
    theme(text = element_text(size = 12))

# plot 6 barplots in one graph
plot_row <- plot_grid(gender_dist, awaken_dist, coffee_dist, alcohol_dist, smoke_dist,
          exercise_dist, nrow = 2)

# now add the title
title <- ggdraw() +
  draw_label(
    "Distributions",
    fontface = 'bold',
    x = 0,
    hjust = 0
  ) +
  theme(
    # add margin on the left of the drawing canvas,
    # so title is aligned with left edge of first plot
    plot.margin = margin(0, 0, 0, 7)
  )
plot_grid(
  title, plot_row,
  ncol = 1,
  # rel_heights values control vertical title margins
  rel_heights = c(0.1, 1)
)
```

Distributions



Here we explore the distributions of all categorical explanatory variables. We see that we have roughly the same number of people for each gender and have the number of nonsmokers twice as large as that of smokers. Awakenings and exercise frequency are more evenly distributed, while caffeine consumption and alcohol consumption are skewed to the right.

```
options(repr.plot.height = 20, repr.plot.width = 20)
sleep_train |>
    ggpairs(columns = c('Age', 'Sleep.duration', 'Sleep.efficiency', 'REM.sleep.percentage', 'Deep.slee
```



Looking at this ggpairs graph which only includes the numerical variables, we see that only deep sleep percentage and light sleep percentage are strongly correlated with the response variable (sleep efficiency). Their correlation are 0.783 and -0.816 respectively.

Almost all the explanatory variables have no significant correlation with each other except deep sleep percentage and light sleep percentage. These two variables have a very strong correlation that is -0.974, and their corresponding scatterplot also exhibits a strong negative linear relationship. This could result in a very problematic issue called collinearity.

Therefore, we will only include deep sleep percentage as one of the explanatory variables.

```
sleep_train <- sleep_train |>
  select(-Light.sleep.percentage)
```

Let's start by making a full model regression which includes all the explanatory variables.

```
full_regression <- lm(Sleep.efficiency ~., data = sleep_train)
full_reg_summ <- summary(full_regression)

full_reg_pred <- predict(full_regression, newdata = sleep_test)
sleep_R_MSE_models <- tibble(
  Model = "Full Regression",
  R_MSE = rmse(
    preds = full_reg_pred,
    actual = sleep_test$Sleep.efficiency,
    na.rm = TRUE),
  AIC = AIC(full_regression),
```

```
  BIC = BIC(full_regression),
  adjusted_R_sq = full_reg_summ$adj.r.squared)
sleep_R_MSE_models
```

```
## # A tibble: 1 x 5
##   Model            R_MSE   AIC   BIC adjusted_R_sq
##   <chr>            <dbl> <dbl> <dbl>         <dbl>
## 1 Full Regression 0.0563 -771. -673.         0.830
```

```
# Here we compute Root Mean Squared Error so that we can use it as a
# comparison metric between best fit models
```

Now we can begin the forward selection process. We also will compute values of RSS, BIC, and Cp to help decide which sized model is the best fit.

```
sleep_forward_selection <- regsubsets(
  x = Sleep.efficiency ~., nvmax = 11,
  data = sleep_train,
  method = "forward")
sleep_forward_selection
```

```
## Subset selection object
## Call: regsubsets.formula(x = Sleep.efficiency ~ ., nvmax = 11, data = sleep_train,
##     method = "forward")
## 25 Variables  (and intercept)
##                          Forced in Forced out
## Age                          FALSE      FALSE
## Gender1                      FALSE      FALSE
## Sleep.duration               FALSE      FALSE
## REM.sleep.percentage         FALSE      FALSE
## Deep.sleep.percentage        FALSE      FALSE
## Awakenings1                  FALSE      FALSE
## Awakenings2                  FALSE      FALSE
## Awakenings3                  FALSE      FALSE
## Awakenings4                  FALSE      FALSE
## Caffeine.consumption25       FALSE      FALSE
## Caffeine.consumption50       FALSE      FALSE
## Caffeine.consumption75       FALSE      FALSE
## Caffeine.consumption100      FALSE      FALSE
## Caffeine.consumption200      FALSE      FALSE
## Alcohol.consumption1         FALSE      FALSE
## Alcohol.consumption2         FALSE      FALSE
## Alcohol.consumption3         FALSE      FALSE
## Alcohol.consumption4         FALSE      FALSE
## Alcohol.consumption5         FALSE      FALSE
## Smoking.status1              FALSE      FALSE
## Exercise.frequency1          FALSE      FALSE
## Exercise.frequency2          FALSE      FALSE
## Exercise.frequency3          FALSE      FALSE
## Exercise.frequency4          FALSE      FALSE
## Exercise.frequency5          FALSE      FALSE
## 1 subsets of each size up to 11
## Selection Algorithm: forward
```

```r
sleep_forward_summary <- summary(sleep_forward_selection)

sleep_forward_summary_df <- tibble(
  n_input_variables = 1:11,
  RSS = sleep_forward_summary$rss,
  BIC = sleep_forward_summary$bic,
  Cp = sleep_forward_summary$cp)

sleep_forward_summary_df
```

```
## # A tibble: 11 x 4
##    n_input_variables   RSS   BIC    Cp
##                <int> <dbl> <dbl> <dbl>
##  1                 1 1.87  -250. 340.
##  2                 2 1.66  -277. 273.
##  3                 3 1.55  -289. 240.
##  4                 4 1.39  -313. 190.
##  5                 5 1.20  -347. 131.
##  6                 6 1.09  -368.  96.4
##  7                 7 0.991 -389.  65.8
##  8                 8 0.945 -396.  52.8
##  9                 9 0.923 -397.  47.7
## 10                10 0.880 -404.  35.8
## 11                11 0.853 -407.  29.1
```

```r
# We can see that the lowest Cp is with 7 variables
summary(sleep_forward_selection)
```

```
## Subset selection object
## Call: regsubsets.formula(x = Sleep.efficiency ~ ., nvmax = 11, data = sleep_train,
##     method = "forward")
## 25 Variables  (and intercept)
##                          Forced in Forced out
## Age                          FALSE      FALSE
## Gender1                      FALSE      FALSE
## Sleep.duration               FALSE      FALSE
## REM.sleep.percentage         FALSE      FALSE
## Deep.sleep.percentage        FALSE      FALSE
## Awakenings1                  FALSE      FALSE
## Awakenings2                  FALSE      FALSE
## Awakenings3                  FALSE      FALSE
## Awakenings4                  FALSE      FALSE
## Caffeine.consumption25       FALSE      FALSE
## Caffeine.consumption50       FALSE      FALSE
## Caffeine.consumption75       FALSE      FALSE
## Caffeine.consumption100      FALSE      FALSE
## Caffeine.consumption200      FALSE      FALSE
## Alcohol.consumption1         FALSE      FALSE
## Alcohol.consumption2         FALSE      FALSE
## Alcohol.consumption3         FALSE      FALSE
## Alcohol.consumption4         FALSE      FALSE
## Alcohol.consumption5         FALSE      FALSE
```

```
## Smoking.status1             FALSE      FALSE
## Exercise.frequency1         FALSE      FALSE
## Exercise.frequency2         FALSE      FALSE
## Exercise.frequency3         FALSE      FALSE
## Exercise.frequency4         FALSE      FALSE
## Exercise.frequency5         FALSE      FALSE
## 1 subsets of each size up to 11
## Selection Algorithm: forward
##           Age Gender1 Sleep.duration REM.sleep.percentage Deep.sleep.percentage
## 1  ( 1 )  " " " "     " "            " "                  "*"
## 2  ( 1 )  " " " "     " "            "*"                  "*"
## 3  ( 1 )  " " " "     " "            "*"                  "*"
## 4  ( 1 )  " " " "     " "            "*"                  "*"
## 5  ( 1 )  " " " "     " "            "*"                  "*"
## 6  ( 1 )  " " " "     " "            "*"                  "*"
## 7  ( 1 )  " " " "     " "            "*"                  "*"
## 8  ( 1 )  "*" " "     " "            "*"                  "*"
## 9  ( 1 )  "*" " "     " "            "*"                  "*"
## 10  ( 1 ) "*" " "     " "            "*"                  "*"
## 11  ( 1 ) "*" " "     " "            "*"                  "*"
##           Awakenings1 Awakenings2 Awakenings3 Awakenings4
## 1  ( 1 )  " "         " "         " "         " "
## 2  ( 1 )  " "         " "         " "         " "
## 3  ( 1 )  " "         " "         "*"         " "
## 4  ( 1 )  " "         " "         "*"         "*"
## 5  ( 1 )  " "         "*"         "*"         "*"
## 6  ( 1 )  "*"         "*"         "*"         "*"
## 7  ( 1 )  "*"         "*"         "*"         "*"
## 8  ( 1 )  "*"         "*"         "*"         "*"
## 9  ( 1 )  "*"         "*"         "*"         "*"
## 10  ( 1 ) "*"         "*"         "*"         "*"
## 11  ( 1 ) "*"         "*"         "*"         "*"
##           Caffeine.consumption25 Caffeine.consumption50 Caffeine.consumption75
## 1  ( 1 )  " "                    " "                    " "
## 2  ( 1 )  " "                    " "                    " "
## 3  ( 1 )  " "                    " "                    " "
## 4  ( 1 )  " "                    " "                    " "
## 5  ( 1 )  " "                    " "                    " "
## 6  ( 1 )  " "                    " "                    " "
## 7  ( 1 )  " "                    " "                    " "
## 8  ( 1 )  " "                    " "                    " "
## 9  ( 1 )  " "                    " "                    " "
## 10  ( 1 ) " "                    "*"                    " "
## 11  ( 1 ) " "                    "*"                    " "
##           Caffeine.consumption100 Caffeine.consumption200 Alcohol.consumption1
## 1  ( 1 )  " "                     " "                     " "
## 2  ( 1 )  " "                     " "                     " "
## 3  ( 1 )  " "                     " "                     " "
## 4  ( 1 )  " "                     " "                     " "
## 5  ( 1 )  " "                     " "                     " "
## 6  ( 1 )  " "                     " "                     " "
## 7  ( 1 )  " "                     " "                     " "
## 8  ( 1 )  " "                     " "                     " "
## 9  ( 1 )  " "                     " "                     " "
```

```
## 10  ( 1 ) " "                     " "                     " "
## 11  ( 1 ) " "                     " "                     " "
##           Alcohol.consumption2 Alcohol.consumption3 Alcohol.consumption4
## 1  ( 1 )  " "                   " "                   " "
## 2  ( 1 )  " "                   " "                   " "
## 3  ( 1 )  " "                   " "                   " "
## 4  ( 1 )  " "                   " "                   " "
## 5  ( 1 )  " "                   " "                   " "
## 6  ( 1 )  " "                   " "                   " "
## 7  ( 1 )  " "                   " "                   " "
## 8  ( 1 )  " "                   " "                   " "
## 9  ( 1 )  " "                   " "                   " "
## 10  ( 1 ) " "                   " "                   " "
## 11  ( 1 ) "*"                   " "                   " "
##           Alcohol.consumption5 Smoking.status1 Exercise.frequency1
## 1  ( 1 )  " "                   " "             " "
## 2  ( 1 )  " "                   " "             " "
## 3  ( 1 )  " "                   " "             " "
## 4  ( 1 )  " "                   " "             " "
## 5  ( 1 )  " "                   " "             " "
## 6  ( 1 )  " "                   " "             " "
## 7  ( 1 )  " "                   "*"             " "
## 8  ( 1 )  " "                   "*"             " "
## 9  ( 1 )  " "                   "*"             "*"
## 10  ( 1 ) " "                   "*"             "*"
## 11  ( 1 ) " "                   "*"             "*"
##           Exercise.frequency2 Exercise.frequency3 Exercise.frequency4
## 1  ( 1 )  " "                 " "                 " "
## 2  ( 1 )  " "                 " "                 " "
## 3  ( 1 )  " "                 " "                 " "
## 4  ( 1 )  " "                 " "                 " "
## 5  ( 1 )  " "                 " "                 " "
## 6  ( 1 )  " "                 " "                 " "
## 7  ( 1 )  " "                 " "                 " "
## 8  ( 1 )  " "                 " "                 " "
## 9  ( 1 )  " "                 " "                 " "
## 10  ( 1 ) " "                 " "                 " "
## 11  ( 1 ) " "                 " "                 " "
##           Exercise.frequency5
## 1  ( 1 )  " "
## 2  ( 1 )  " "
## 3  ( 1 )  " "
## 4  ( 1 )  " "
## 5  ( 1 )  " "
## 6  ( 1 )  " "
## 7  ( 1 )  " "
## 8  ( 1 )  " "
## 9  ( 1 )  " "
## 10  ( 1 ) " "
## 11  ( 1 ) " "
```

We found that the best model has 7 variables and found which ones from the chart above. Now we can put that into a model and compare it with the original.

```r
#make the model with 7 variables
sleep_selected_regression <- lm(Sleep.efficiency ~ Age + REM.sleep.percentage +
                                  Deep.sleep.percentage + Awakenings +
                                  Alcohol.consumption + Smoking.status +
                                  Exercise.frequency, data = sleep_train)
selected_reg_summ <- summary(sleep_selected_regression)

#fit a predicted model
sleep_selected_reg_pred <- predict(sleep_selected_regression, newdata = sleep_test)

#Now we can compare the RMSE on the models
sleep_R_MSE_models <- rbind(
  sleep_R_MSE_models,
  tibble(
    Model = "Selected Regression",
    R_MSE = rmse(
      preds = sleep_selected_reg_pred,
      actuals = sleep_test$Sleep.efficiency,
      na.rm = TRUE),
  AIC = AIC(sleep_selected_regression),
  BIC = BIC(sleep_selected_regression),
  adjusted_R_sq = selected_reg_summ$adj.r.squared))
sleep_R_MSE_models
```

```
## # A tibble: 2 x 5
##   Model                R_MSE   AIC   BIC adjusted_R_sq
##   <chr>                <dbl> <dbl> <dbl>         <dbl>
## 1 Full Regression     0.0563 -771. -673.         0.830
## 2 Selected Regression 0.0556 -802. -729.         0.817
```

We also tried to add some interaction terms to explore some potential relationship: Here we create interactions of age and other variables and split them into 4 different models. The reason why we choose Age is that there are a lot of factors may be affected by age. For example, the damage of smoking may vary for people in different ages.

```r
#model with an interaction term Age*Smoking.status
sleep_inter1 <- lm(Sleep.efficiency ~ Age * Smoking.status +
                                REM.sleep.percentage + Awakenings +
                                Alcohol.consumption + Deep.sleep.percentage +
                                Exercise.frequency, data = sleep_train)

sleep_inter1_summ <- summary(sleep_inter1)

sleep_inter1_pred <- predict(sleep_inter1, newdata = sleep_test)

sleep_R_MSE_models <- rbind(
  sleep_R_MSE_models,
  tibble(
    Model = "Selected Regression with Interaction Age*Smoking.status",
    R_MSE = rmse(
      preds = sleep_inter1_pred,
      actuals = sleep_test$Sleep.efficiency,
      na.rm = TRUE),
```

```
  AIC = AIC(sleep_inter1),
  BIC = BIC(sleep_inter1),
  adjusted_R_sq = sleep_inter1_summ$adj.r.squared))



#model with an interaction term Age*REM.sleep.percentage
sleep_inter2 <- lm(Sleep.efficiency ~ Age * REM.sleep.percentage + Smoking.status +
                             Awakenings + Alcohol.consumption + Deep.sleep.percentage +
                              Exercise.frequency, data = sleep_train)

sleep_inter2_summ <- summary(sleep_inter2)

sleep_inter2_pred <- predict(sleep_inter2, newdata = sleep_test)

sleep_R_MSE_models <- rbind(
  sleep_R_MSE_models,
  tibble(
    Model = "Selected Regression with Interaction Age*REM.sleep.percentage",
    R_MSE = rmse(
      preds = sleep_inter2_pred,
      actuals = sleep_test$Sleep.efficiency,
      na.rm = TRUE),
  AIC = AIC(sleep_inter2),
  BIC = BIC(sleep_inter2),
  adjusted_R_sq = sleep_inter2_summ$adj.r.squared))




#model with an interaction term Age*Deep.sleep.percentage
sleep_inter3 <- lm(Sleep.efficiency ~ Age * Deep.sleep.percentage + REM.sleep.percentage +
                     Smoking.status + Awakenings + Alcohol.consumption +
                    Exercise.frequency, data = sleep_train)

sleep_inter3_summ <- summary(sleep_inter3)

sleep_inter3_pred <- predict(sleep_inter3, newdata = sleep_test)

sleep_R_MSE_models <- rbind(
  sleep_R_MSE_models,
  tibble(
    Model = "Selected Regression with Interaction Age*Deep.sleep.percentage",
    R_MSE = rmse(
      preds = sleep_inter3_pred,
      actuals = sleep_test$Sleep.efficiency,
      na.rm = TRUE),
  AIC = AIC(sleep_inter3),
  BIC = BIC(sleep_inter3),
  adjusted_R_sq = sleep_inter3_summ$adj.r.squared))
```

```
#model with an interaction term Age*Awakenings
sleep_inter4 <- lm(Sleep.efficiency ~ Age * Awakenings + Deep.sleep.percentage +
                     REM.sleep.percentage + Smoking.status + Alcohol.consumption +
                    Exercise.frequency, data = sleep_train)

sleep_inter4_summ <- summary(sleep_inter4)

sleep_inter4_pred <- predict(sleep_inter4, newdata = sleep_test)

sleep_R_MSE_models <- rbind(
  sleep_R_MSE_models,
  tibble(
    Model = "Selected Regression with Interaction Age*Awakenings",
    R_MSE = rmse(
      preds = sleep_inter4_pred,
      actuals = sleep_test$Sleep.efficiency,
      na.rm = TRUE),
  AIC = AIC(sleep_inter4),
  BIC = BIC(sleep_inter4),
  adjusted_R_sq = sleep_inter4_summ$adj.r.squared))
sleep_R_MSE_models
```

```
## # A tibble: 6 x 5
##   Model                                                R_MSE   AIC   BIC adjus~1
##   <chr>                                                <dbl> <dbl> <dbl>   <dbl>
## 1 Full Regression                                     0.0563 -771. -673.   0.830
## 2 Selected Regression                                 0.0556 -802. -729.   0.817
## 3 Selected Regression with Interaction Age*Smoking.s~ 0.0559 -803. -727.   0.819
## 4 Selected Regression with Interaction Age*REM.sleep~ 0.0558 -801. -724.   0.817
## 5 Selected Regression with Interaction Age*Deep.slee~ 0.0536 -805. -728.   0.820
## 6 Selected Regression with Interaction Age*Awakenings 0.0547 -795. -707.   0.815
## # ... with abbreviated variable name 1: adjusted_R_sq
```

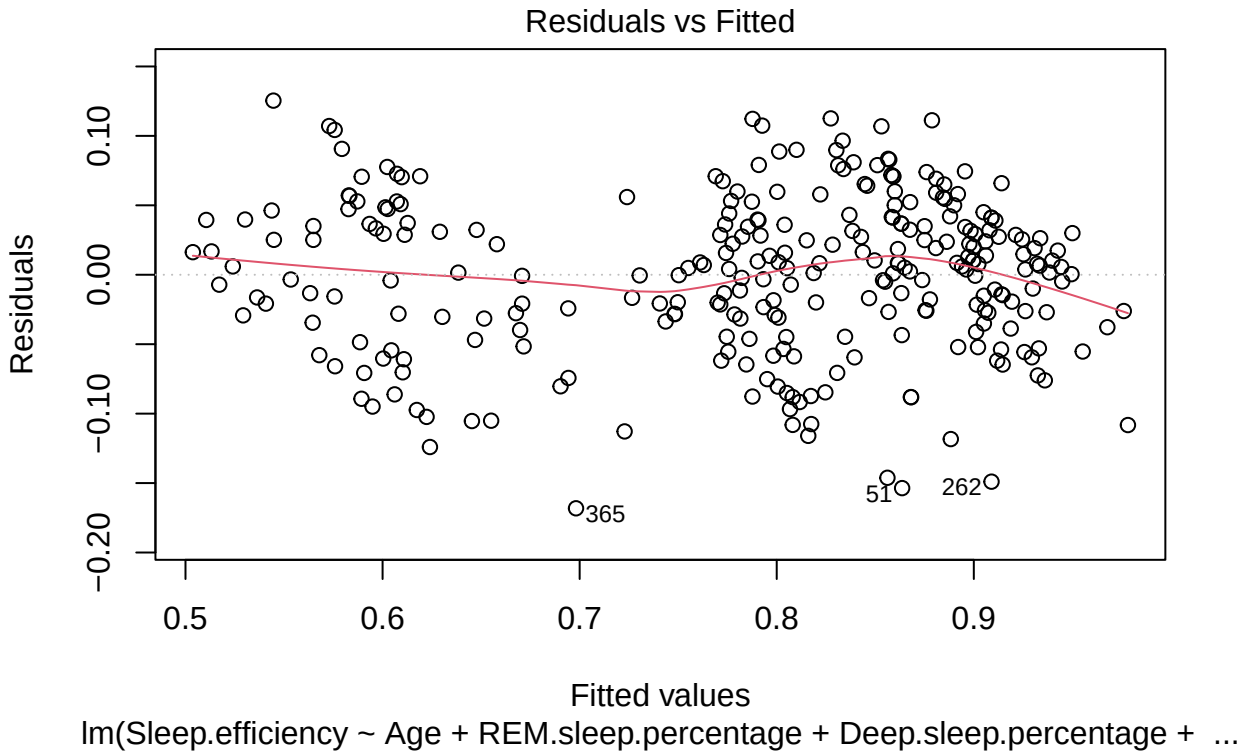
From here, we can find that the model with the lowest RMSE value, lowest AIC value, lowest BIC value, and highest adjusted R squared value gives the best out-of-sample prediction performance.

In this example, we can see that the regression model where we selected the variables using forward regression produces a better model. The six model we build all have roughly the same value of RMSE and adjusted R square. The full regression model has slightly higher RMSE, but it also has slightly higher adjusted R square value. Therefore, we will use AIC and BIC value to compare models. Here all of the selected regression models (with or without interaction terms) have way smaller AIC and BIC values than the full regression model, but the values between each selected model do not differ much (AIC = -800, BIC = -720 approximately). Selected Regression with Interaction Age*Deep.sleep.percentage has the lowest AIC, and Selected Regression without interaction term has the lowest BIC value. Since models with less number of parameters are preferred, the Selected Regression model without interaction term is the best model.

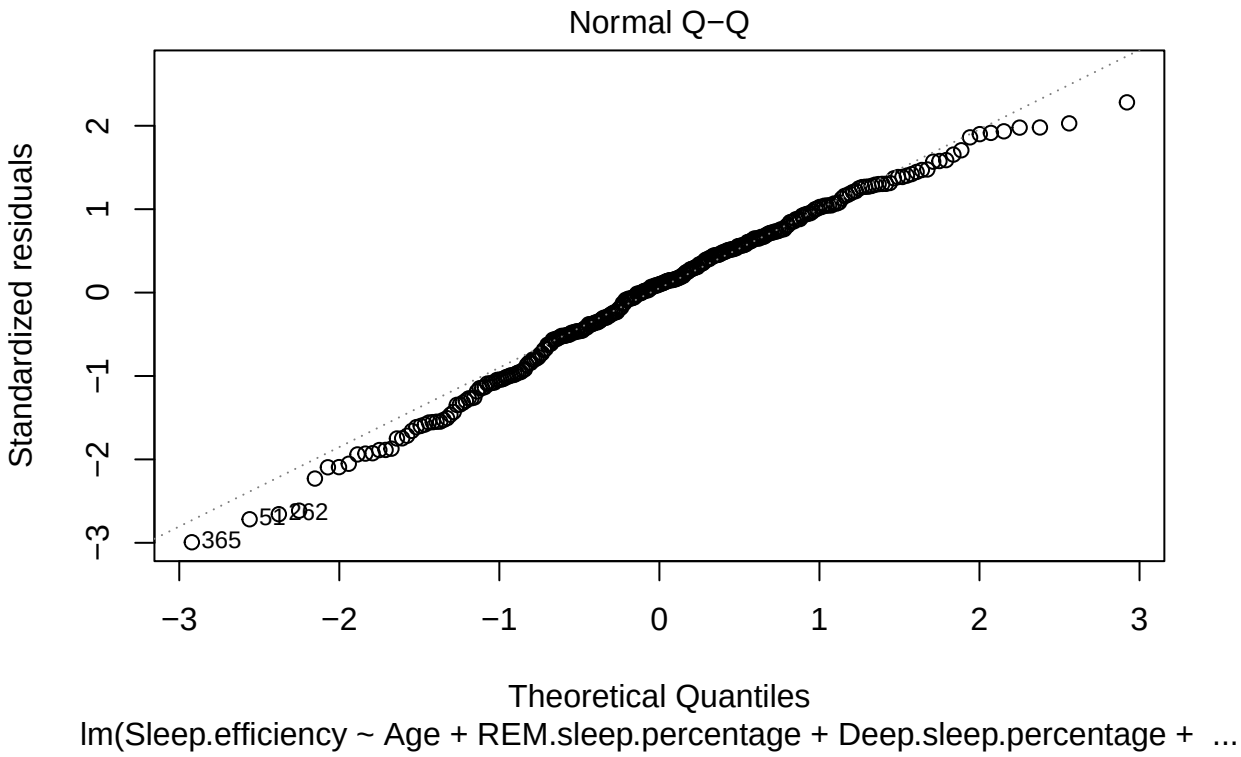
Based on the conclusion above, we showed that selecting variables helps improve the model, in which we used a forward selection process to find that. We also found that in a model with 7 out of 11 selected variables: Gender, Sleep duration, and Caffeine Consumption were not as helpful in comparison to the other variables when predicting Sleep efficiency- which is an interesting discovery. It helps us to answer our research question which asks which variables are most important. We were surprised to find that caffeine consumption wasn't as important to be kept in the model, as we found many studies when researching that implied the opposite. This could be explained by that we only have caffeine consumption as a categorical variable with 4 levels in our model, and hence it could be difficult to fit a linear regression on this limited data we have. If we

were able to collect more data (not 4 levels, but a continuous variable precise to 2 decimal places how much caffeine people consume in 24 hours), then we could treat caffeine consumption as a numerical variable and fit a more accurate linear regression model based on it.

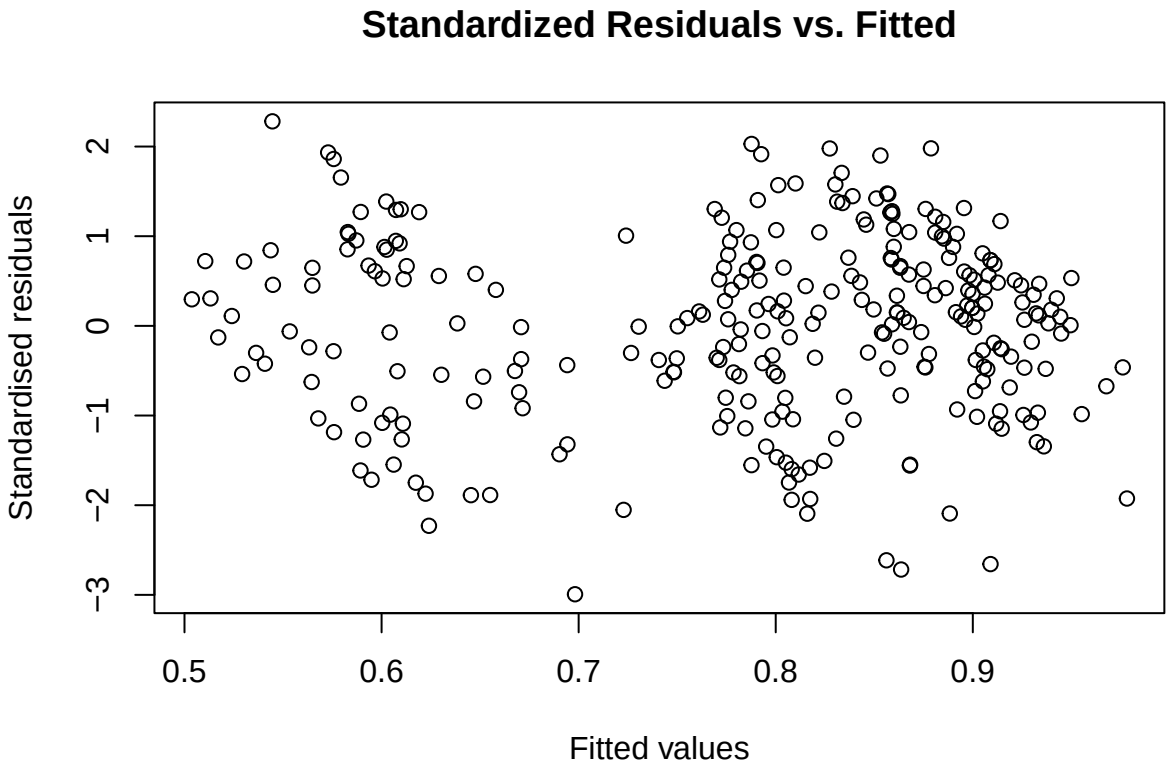
```
plot(sleep_selected_regression, 1)
```




```
plot(sleep_selected_regression, 2)
```




```
plot(fitted(sleep_selected_regression), rstandard(sleep_selected_regression),
     main = 'Standardized Residuals vs. Fitted',
     xlab = 'Fitted values', ylab = 'Standardised residuals')
```



Here we see that the best model we choose may not be the most suitable. It seems that the residual plot does not have a constant variance, and the data points are clustered in 2 groups, but the residuals follows a Normal distribution approximately. The standardized residual plot also shows that we may have a problem with our model since there are too many data points lying outside the range [-2, 2]. If they follow the Normal distribution exactly, there should be only 3 data points outside this range given the empirical rule.

Therefore, we can still improve our model perhaps by collecting more data and include more variables especially changing some categorical variables into numerical.

#In summary, based on our linear regression model, we find that sleep efficency is closely related to people's age, REM.sleep.percentage, Deep.sleep.percentage, awakening times and Smoking.status. #Before the research started, we expected the alcohol comsumption, exercise frequency and caffine comsumption would be the most important factors, but the result shows differet. However the age #is also a factor that will increase the sleep efficiency is what we unexpected. The other exploratory variables come as no surprise, smoking and awakening will decrease you sleep efficency, and deep #sleep and REM sleep will give you a high quality sleep.